Data Mining and Information Retrieval in the 21st century: A bibliographic review

Article *in* Computer Science Review · November 2019

DOI: 10.1016/j.cosrev.2019.100193

Xiangjie Kong
Zhejiang University of Technology
138 PUBLICATIONS **2,226** CITATIONS

Ivan Lee
University of South Australia
134 PUBLICATIONS **1,207** CITATIONS

Feng Xia
Federation University Australia
382 PUBLICATIONS **7,850** CITATIONS

Data Mining and Information Retrieval in the 21st Century: A Bibliographic Review

Jiaying Liu[a], Xiangjie Kong[a,*], Xinyu Zhou[a], Lei Wang[a], Da Zhang[b], Ivan Lee[c], Bo Xu[a], Feng Xia[a]

[a] *Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China.*

[b] *Electrical and Computer Engineering, University of Miami, 5452 Coral Gables, Miami, FL, 33124, USA.*

[c] *School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, Australia.*

Abstract

Data Mining and Information Retrieval is an emerging interdisciplinary discipline dealing with Information Retrieval and Data Mining techniques. It has undergone rapid development with the advances in mathematics, statistics, information science, and computer science. In this paper, we present an empirical analysis of publication metadata obtained from 6 top-tier journals and 9 conferences for the first 16 years of the 21st Century, and evaluate the dynamic characteristics of Data Mining and Information Retrieval. We find a steady growth both in terms of productivity and impact, evidenced by the unabated number of publications/citations over the period of study. We note that the modality for co-operation in this field is changing from independent to collaborative. Furthermore, according to the citation pattern, the field is becoming open-minded as illustrated by a gradual decline of self-citation rates, which was dropped to 10% in 2015, nearly three times lower than what it was in 2000. Finally, we explore the inner structure relying on the topics evolution from the aspects of popular keywords/topics identification and evolution. Overall, this study provides insights of Data Mining and Information Retrieval behind its demonstrated growth in the recent past, with the ultimate goal of revealing its

[*] Corresponding author; email: xjkong@ieee.org.

potential of driving scientific innovation in the future.

Keywords: Data Mining and Information Retrieval, scientific innovation, data analytics, Science of Science

2010 MSC: 00-01, 99-00

1. Introduction

Data Mining and Information Retrieval is coupling of scientific discovery and practice, whose subject is to collect, manage, process, analyze, and visualize the vast amount of structured or unstructured data. It has grown dramatically and became more institutionalized in the 21^{st} Century. Actually, it is closely related to data statistics, whose subject is "learning from data". Data mining refers to the process of searching hidden information from a large number of data through algorithms [1]. Information Retrieval covers algorithms dealing with retrieval subsets from the large collections based on users' need [2]. So they employ multi-disciplinary studies from mathematics, statistics, information science, and computer science, by utilizing techniques such as machine learning, classification, cluster analysis, data mining, databases, and visualization.

On the grounds of exploring the anatomy of scholarly big data [3, 4, 5], and Science of Science [6, 7], including scientific collaboration [8, 9, 10, 11], scientific evolution and recommendation [12, 13], and homogeneous structure [14], researchers have better understandings of the inner structure of the field [15] and how it has evolved over time. In recent years, there has been an increasing amount of literature on emphasizing the ability of quantitative analysis to obtain the overview of a specific field. Sun et al. [16] built a co-authorship network using publication metadata from 22 transportation journals to understand scientific collaborations in transportation research. Meyer et al. [17] measured the development of social simulation using citation and co-citation analysis focusing on the scientific publication. Sinatra et al. [18] analyzed the Web of Science data spanning more than a century to reveal the rapid growth of physics. Iqbal et al. [19] provides a comprehensive bibliometric analysis in the field of computer networking from the perspectives of metadata analysis, content-based analysis, and citation analysis. Based on the large amount of publication metadata [20, 21, 22], these scientometric exercises are valuable to give an overview of the specific field of research and help researchers understand the evolution of the field.

Although the analysis of specific research field attracts more and more attention [23, 24, 25, 26, 23], few of them focus on Data Mining and Information Retrieval due to the difficulty in obtaining its dynamic character. The dynamic nature of this field remains unclear. In this paper, we investigate the dynamic development to understand the field of Data Mining and Information Retrieval at the beginning of the 21^{st} Century. We use the publication metadata obtained from 6 top-tier scientific journals and 9 conferences. This study examines the dynamic nature of Data Mining and Information Retrieval to better identify, quantify, and understand the evolution and trends over 2000-2015 from following aspects: The growth of Data Mining and Information Retrieval, impact and citation pattern analysis, identifying important papers/researchers/institutions, and inner structure exploration based on topics evolution. Our paper aims at providing the guidance for researchers, institutions, funding agencies to make informed decisions.

The remaining part of the paper is structured as follows: Section 2 is concerned with the methodology used for this study, including the introduction of the publication metadata used throughout this paper, and various measures to quantify the development by aggregating the result at levels of the number of publications, citations including self-citations, and time. Section 3 presents the findings of the research, focusing on four essential themes: The growth of Data Mining and Information Retrieval, the dynamic of citation pattern and impact, identification of the influential papers/researchers/institutions, and historical evolution of topics in the 21^{st} Century. Finally, the study is summarized and the future work is highlighted in Section 4.

2. Methodology

2.1. Dataset

To give a deep insight into the science of Data Mining and Information Retrieval in the 21^{st} Century, we use the large-scale scholarly dataset sourced from Microsoft Academic Graph (MAG)[1], which is provided by Microsoft Academic Services [27]. The MAG dataset contains more than 100 million publication metadata. We focus on the typical publications published in the representative journals/conferences during 2000-2015. It is essential for the study to detect the Data Mining and Information Retrieval papers. Here we regard the paper published in the Data Mining and Information Retrieval journals as a Data mining and information retrieval paper because it is easy for us to profile the area. Hence, we select top-tier journals and conferences both in the list of China Computer Federation (CCF)[2] recommended international academic publications and Computing Research and Education Association of Australasia (CORE)[3], which provide recognized rankings under the categories "Database/Data Mining/Information Retrieval" and "Data Science and Engineering".

Furthermore, in this paper, the analysis is distinguished by presenting two sets of data results from journals data or proceedings data. There are two main reasons for taking into the difference based on the type of documents:

1. **Growth of the publications/citations**. Generally speaking, the conferences are held every one to two years. In contrast, journals are usually published monthly or quarterly. Different publication time may lead to fluctuations in the growth of publications/citations.
2. **Topic detection**. As we all know, it takes about 2 to 4 months for a conference paper from submission to acceptance. At the same time, journal papers need to be reviewed and revised many times, which will

[1] https://www.openacademic.ai/oag/
[2] https://www.ccf.org.cn/xspj/gyml
[3] http://www.core.edu.au

take even one or two years. It is possible that hot topics in the same year for journal papers and conference papers are totally different.

Table 1 and Table 2 provide the specific information about these journals and conferences, including the total number of papers published between 2000 and 2015, citations of these papers, and the total number of unique authors, respectively. Based on these general data, we calculate the average number of authors per paper, the average number of published papers per author, and the average number of citations per paper. The average number of authors per paper represents the average length of the author list, which is computed as $\sum_{p \in P} |au_p|/|P|$, and it can be used to measure the degree of collaboration. Similarly, the average number of papers per author on behalf of the average productivity, which is computed as $|P|/\sum_{p \in P} |au_p|$. In these formulas, $|P|$ represents the total number of papers published in the journal/conference, and $|au_p|$ is the total number of authors in these publications. Note that, for the computation of the average number of papers, the au_p is unique authors.

Table 1: Statistics for each journal.

Journal name	Papers	Authors	Citations	Authors per paper	Authors per paper (max)	Papers per author	Citations per paper
European Journal of Information Systems	833	1353	16131	2.12	10	0.62	19.36
IEEE Transactions on Knowledge and Data Engineering	2464	5057	55603	3.11	10	0.49	22.57
Journal of the Association for Information Science and Technology	212	345	3426	1.87	7	0.61	16.16
ACM Transactions on Database Systems	402	858	12033	3.09	10	0.47	29.93
ACM Transactions on Information Systems	353	876	15526	3.16	11	0.40	43.98
The VLDB Journal	1809	3587	62809	3.46	27	0.50	34.72

Table 2: Statistics for each conference.

Conference name	Papers	Authors	Citations	Authors per paper	Authors per paper (max)	Papers per author	Citations per paper
ACM International Conference on Web Search and Data Mining (WSDM)	539	1239	10155	3.34	12	0.44	18.84
ACM SIGMOD/PODS International Conference on Management of Data (SIGMOD)	2487	4832	68469	3.66	30	0.51	27.53
IEEE International Conference on Data Engineering (ICDE)	3422	6758	45725	3.44	36	0.51	13.36
International World Wide Web Conferences (WWW)	4665	10168	91200	3.32	24	0.46	19.55
IEEE International Conference on Data Mining series (ICDM)	3981	8536	30568	3.17	14	0.47	7.68
ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)	3915	7988	85584	3.29	34	0.49	21.86
Symposium on Principles of Database Systems (PODS)	526	724	14219	2.65	8	0.73	27.03
International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)	3112	4651	57920	3.04	28	0.67	18.61
International Conference on Very Large Data Bases (VLDB)	2619	5078	76298	3.64	27	0.52	29.13

The average number of citations per paper ($\sum_{p\in P}|ci_p|/|P|$) for each journal/conference enables us to calculate the impact of the journal/conference, where $|P|$ is the total number of publications published in the specific journal/conference, and $|ci_p|$ is the total number of citations these publications received.

2.2. Self-citation rate

Studies have emphasized the ability of self-citation analysis to understand the nature of academic network [28, 29], which can reveal the referencing behavior of scientists. In this paper, we consider two kinds of self-citation: author self-citation, journal self-citation, and conference self-citation.

Author self-citation, the citation relationship between two publications with the same author, can be regarded as the major way in which researchers can seek to rhetorically construct their professional credibility. We use the most rigorous evaluation criteria. We define that if one paper cites another paper with at least one common author, then the citation is an author-self citation. According to this definition, the author self-citation rate (ASR) of a paper means the proportion of author self-citations in total citations.

Definition 1. *ASR is defined as the author self-citation rate of a paper, i.e.,*

$$ASR = \sum_{c \in C} |ac_c| / |C|$$

where $|C|$ is the total number of citations a paper received, and ac_c is the author self-citation.

Journal self-citation is also a crucial indicator which is widely used to measure the contribution of a journal. It represents that the cited paper and the citing paper are published in the same journal. The self and external influence of journals can be calculated in terms of journal self-citation rate. Similarly, we define the conference self-citation to measure the external influence of conferences.

Definition 2. *JSR is defined as the self-citation rate of a journal, i.e.,*

$$JSR = \sum_{c \in C} |jc_c| / |C|$$

where $|C|$ is the total number of citations a journal received, and jc_c is the journal self-citation.

Definition 3. *CSR is defined as the self-citation rate of a conference, i.e.,*

$$CSR = \sum_{c \in C} |cc_c| / |C|$$

where $|C|$ is the total number of citations a conference received, and cc_c is the conference self-citation.

2.3. Topics discovery in Data Mining and Information Retrieval

To give a deep insight into the inner structure of this multi-disciplinary field, we recognize the topics for each paper according to the field of study hierarchy.

MAG dataset provides the studyIDs based on the keywords of the paper. It also provides the child field and parent field of each studyID. By using the information, we build a hierarchy tree for these study fields and serve the second-order parent as topics within the field of Data mining and information retrieval. In order to obtain the number of publications and the citation relationship for these topics, the steps are described as follows:

1. We identify the keywords for each paper, and match the field of studyID to keywords in the dataset;
2. According to hierarchies provided by MAG, we build a hierarchy tree for each studyID, take the studyID (L_3) as the root, and mark it's parent as L_2. After this process, the parent of L_2 is L_1 and the top of the tree is marked as L_0;
3. For each studyID (L_3), we find its second-order parent L_1, and accept it as the topic;
4. We map the paper as well as the citation relationship to topics;
5. We calculate the number of publications and citations for each topic.

2.4. Development index and increase index of topics

To reflect the evolution of topics, here we define two indices: development index (DI) and increase index (II).

- DI: The proportion of the topic during specific time. It is calculated by the proportion of papers that contain the specific topic in all papers.

- II: The increase index between two time periods for each topic: 2000-2007 and 2008-2015. It is calculated based on DI_k for topic k. The computational formula is $\sum_{t=2008}^{2015} DI_k / \sum_{t=2000}^{2007} DI_k$.

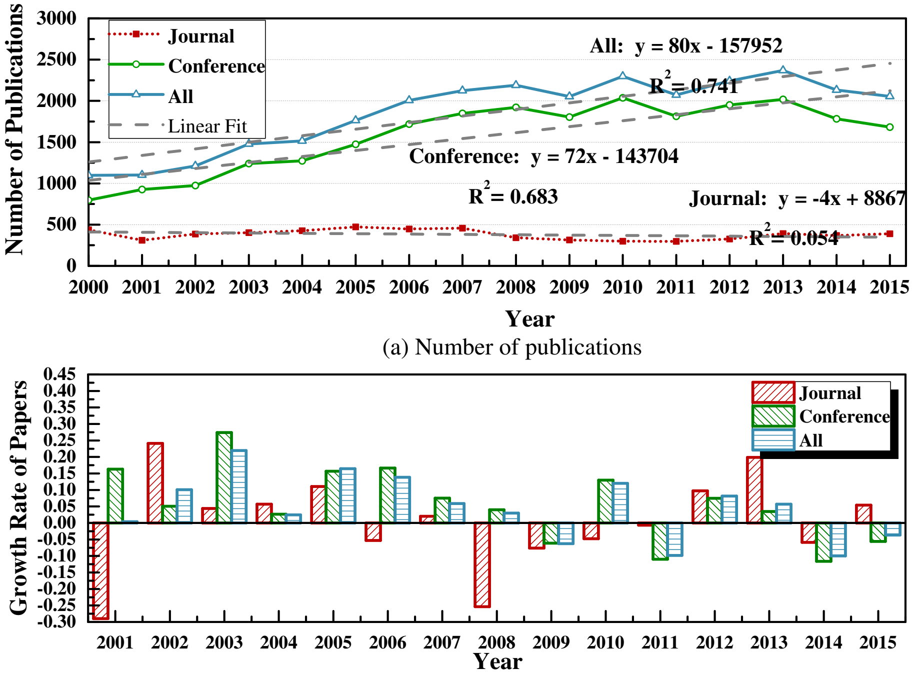
3. Result

3.1. The growth of Data Mining and Information Retrieval

The development history of Data Mining and Information Retrieval, such as the renewal of scientific data research methodology and data representation methodology, leads to a large number of publications. Data Mining and Information Retrieval as an application science, combining with other fields, derive various interdisciplinary fields, such as behavioral Data mining and information retrieval, brain data science, meteorology data science, financial data science, geography data science and so on, whose continuous development greatly promoted the progress of Science. The increasing number of publications each year serves as an evidence in support of this.

Figure 1(a) shows that the number of Data Mining and Information Retrieval papers keeps increasing. However, the growth rate of journal papers does not show an obvious upward trend. There are two possible reasons for the phenomenon. On the one hand, the number of papers published in the journals is roughly the same every year. On the other hand, generally, compared with journals, conferences are held to show the most advanced ideas and technologies. Simultaneously, conferences provide forums for scientists and researchers to communicate and exchange ideas with those from other institutions. They can publish their achievements in a short period of time, which is favorable for subjects as their findings require timely dissemination. In contrast, journals require more complete results, which may result in the fluctuation of the growth rate. So the growth rate of conference papers is generally higher than journal papers (see in Figure 1(b)).

To explore the reasons for the increase in the number of papers in-depth, we study the growth of authors (Figure 2(a)) and find that the growth rate (Figure 2(b)) also has the upward trend but higher than the number of publications. We can also see the number of authors does not show a modest trend and keeps growing fast, which means the heat of Data Mining and Information Retrieval is undiminished. More and more scientists are delving to this field. Therefore, we conclude that the growth of Data Mining and Information Retrieval is partially driven by the increasing number of authors. Figure 2(c) shows that the average number of authors per paper keeps increasing over time which indicates that cooperation is becoming more and more common. Figure 2(d) illustrates that the average number of publications per author has a declining trend from 0.55

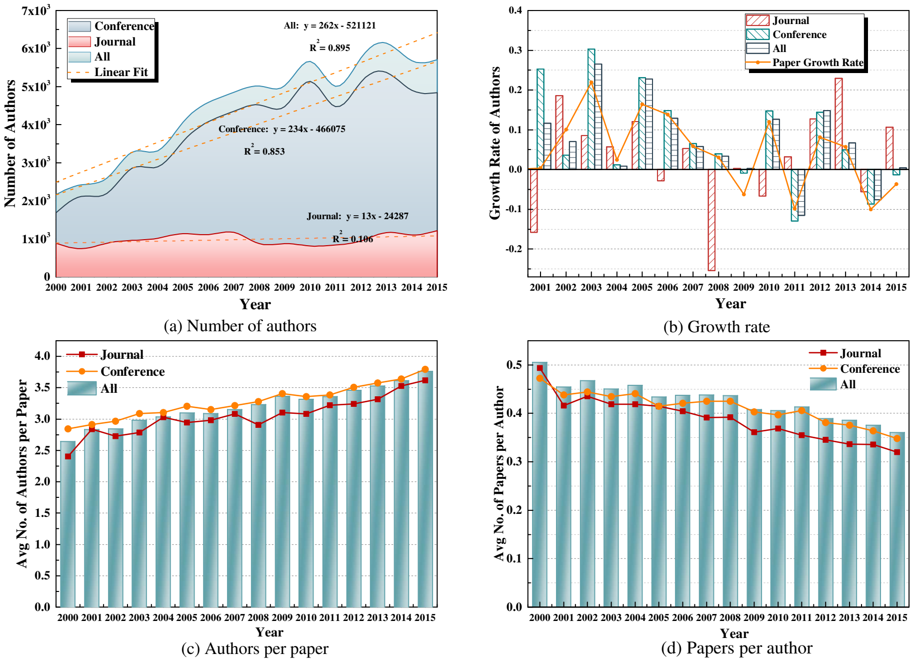


(a) Number of publications

(b) Growth rate of papers

Figure 1: The evolution of the number of publications in the 21^{st} Century for the area of Data Mining and Information Retrieval. (a) The number of publications each year. The gray dotted lines are their respective fitting lines. (b) The growth rate of publications over time.

to 0.45 during 2000 to 2012. After 2012, the average productivity still follows a decline trend with slower speed and decrease to 0.35 till 2015 zigzaggedly. This phenomenon is caused not only by the increasing number of authors but also originated from the decreasing speed of development of Data Mining and Information Retrieval in the 21st Century. The production decrease maybe since authors not only publish in the selected set of journals.



(a) Number of authors

(b) Growth rate

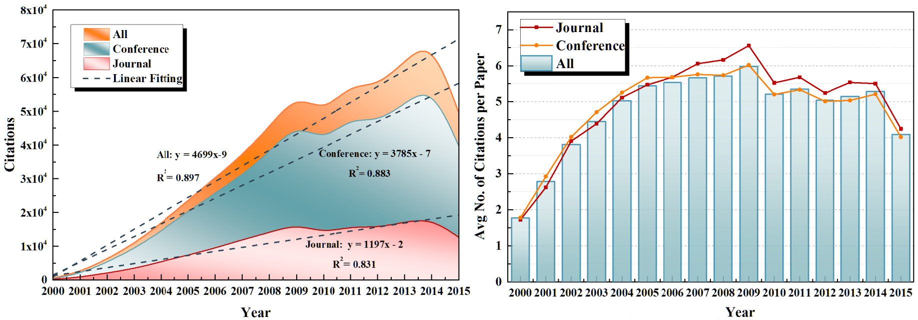
(c) Authors per paper

(d) Papers per author

Figure 2: Number of authors over time in the field of Data Mining and Information Retrieval. (a) The number of authors each year. The orange dotted lines are fitting lines. (b) The author growth rate as well as paper growth rate each year. The histogram is the growth rate of number of authors and the line represents the growth rate of all publications. (c) The average number of authors per paper. (d) The average productivity in terms of number of publications per author over time.

3.2. Impact and citation pattern analysis

While comparing citation numbers across years may be unfair [30], we use both the citation count and the average number of citations per paper to explore the impact of Data Mining and Information Retrieval. Comparing Figure 1(a) with Figure 3(a), we can see that citations increase in a much faster pace than the number of publications. Figure 3(b) shows that the average number of citations per paper has been increasing continuously till 2008, which reflects the increasing impact of Data Mining and Information Retrieval to other areas. After 2008, it fluctuates and decreases, which is related to the late publications' year. The sharp growth of citations can be explained from two aspects: the increasing number of references per paper and the increasing number of publications.

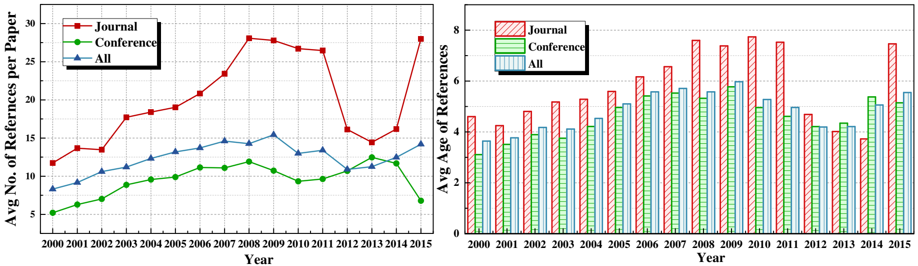


(a) Changes in the citations

(b) Average of citations per paper

Figure 3: The evolution of the influence of Data Mining and Information Retrieval. (a) The number of total citations of publications in Data Mining and Information Retrieval each year. The grey dotted lines are their respective fitting lines. (b) The average number of citations per paper each year.

We then analyze the average number of references per paper from 2000 to 2015, and present results in Figure 4(a). The average number of references per journal paper rises ceaselessly from 13 in 2000 to 28 in 2008, conferences have the same trend from 6 in 2000 to 13 in 2009. Overall, the length of the reference list for conference papers is lower than that of journal papers. As mentioned above, conference papers pay more attention to the ideas and early stage findings for discussion, as oppose to completeness of the study as required in typical journals. Thus, journal papers discover deeper and cover wider in certain research area, which may increase the amount of references. Figure 4(b) shows a periodic change in the average reference age of the paper, which also suggests that the results need a periodic time from publication to actual application [31].



(a) Average number of references

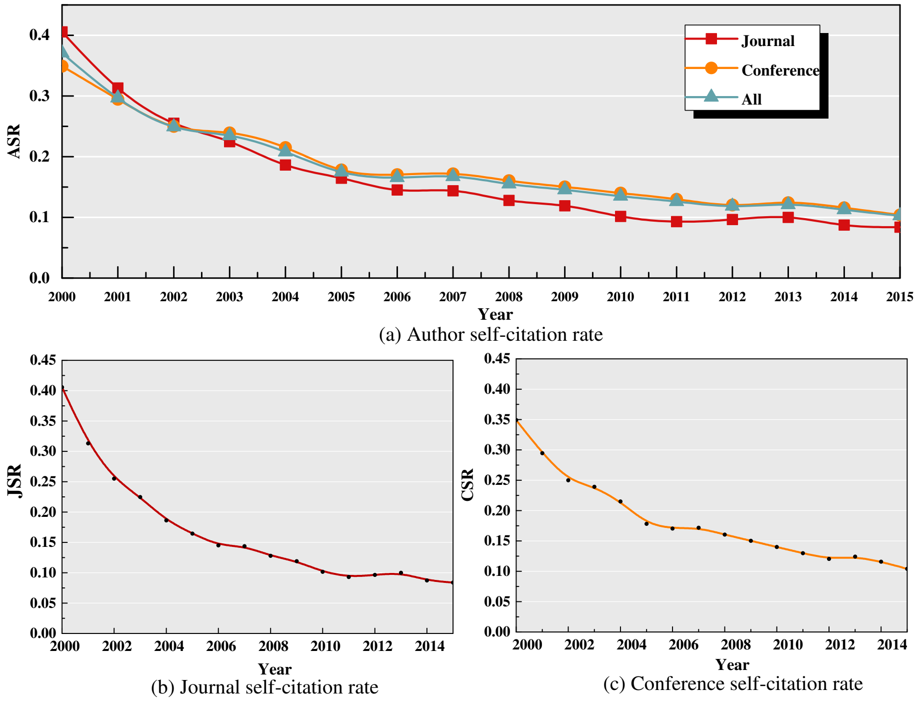
(b) Average age of references

Figure 4: The evolution of references of the publications in Data Mining and Information Retrieval. (a) The average number of references per paper each year. (b) The average age of references.

According to Figure 5(a), we observe that author self-citation decreases from about 0.38 in 2000 to 0.1 in 2015, which indicates that researchers' influence is expanding. Figure 5(b) and Figure 5(c) show the self-citation of journals and conferences. They both reduced from about 0.4 to 0.1 and the trend continues to decline, which means Data Mining and Information Retrieval persists sustainable development and has attracted more attention from other fields.

3.3. Identification of important papers/researchers/institutions

In this subsection, we apply the introduced measures in 2 to quantify the importance of papers/researchers/institutions in the era of Data Mining and Information Retrieval. For papers, we use the total number of citations it received during 2000-2015 to measure their impact. The papers are divided into journal papers and conference papers. From Table 3, we can identify essential topics and the change of hot topics in the Data Mining and Information Retrieval's development. Their main topics range from recommendation systems (algorithms) and databases to data stream, sensor network, and social network. It's not difficult to conclude that the number of citations heavily rely on the type of topics addressed by the community. The study [32] has already suggested that the number of citations received per year shows a typical birth-death process. Here, "Birth-Death" process means that papers receive few citations at the beginning



(a) Author self-citation rate

(b) Journal self-citation rate

(c) Conference self-citation rate

Figure 5: The evolution of self-citation behavior. (a) The rate of author self-citation (ASR).(b) The average journal self-citation rate (JSR) of the top journals each year. (c) The conference self-citation rate (CSR) of the top conferences over time.

and more citations later. After that, the number of citations diminishes due to the obsolete content. The popularity of the topic can be one of the important factors in determining the impact of the paper. These important papers obtain lots of attention during this time period.

After identifying important papers, we focus on measuring researchers' impact. We compute statistics of the total number of publications in the top conferences/journals for each researcher during 2000-2015 and the citations these publications received. After that, we calculate the average citation per paper for these researchers, and then rank them based on results. These influential researchers are located at the tails of distributions and have widespread reputation. The significant correlation can be found between the ranking of papers and the ranking of researchers. For example, Reutemann and Hall [33] partici-

Table 3: Ranking of publications ordered by the total number of citations received between 2000 and 2015.

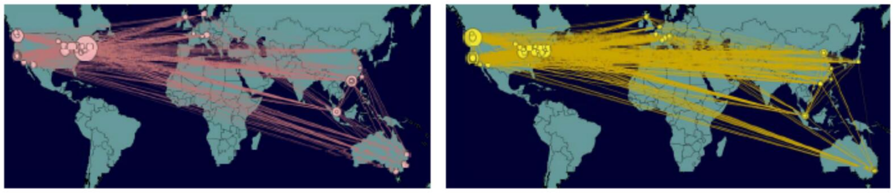
(a) Journal

No.	Title	Citations	Published year
1	Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions	2487	2005
2	A survey of approaches to automatic schema matching	1799	2001
3	Evaluating collaborative filtering recommender systems	1736	2004
4	Cumulated gain-based evaluation of IR techniques	920	2002
5	TinyDB: An acquisitional query processing system for sensor networks	891	2005
6	Protecting respondents identities in microdata release	810	2001
7	A survey on transfer learning	800	2010
8	Toward integrating feature selection algorithms for classification and clustering	709	2005
9	PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities	667	2004
10	Duplicate record detection: A Survey	667	2007
11	Workflow mining: Discovering process models from event logs	619	2004
12	Learning from imbalanced data	600	2009
13	A framework for clustering evolving data streams	590	2003
14	Answering queries using views: A survey	578	2001
15	Item-based top-N recommendation algorithms	557	2004
16	Measuring praise and criticism: Inference of semantic orientation from association	556	2003
17	Graph clustering by flow simulation	545	2001
18	Latent semantic models for collaborative filtering	544	2004
19	Model-driven data acquisition in sensor networks	515	2004
20	RoadRunner: Towards automatic data extraction from large web sites	508	2001
21	Combating web spam with trustrank	506	2004
22	Indexing and querying XML data for regular path expressions	502	2001
23	A study of smoothing methods for language models applied to information retrieval	492	2004
24	Scalable algorithms for association mining	487	2000
25	COMA: A system for flexible combination of schema matching approaches	472	2002
26	The Google similarity distance	466	2007
27	Doing interpretive research	452	2006
28	Efficient query evaluation on probabilistic databases	450	2004
29	Approximate frequency counts over data streams	450	2002
30	The CQL continuous query language: Semantic foundations and query execution	438	2006

(b) Conference

No.	Title	Citations	Published year
1	The WEKA data mining software: An update	4592	2009
2	Mining frequent patterns without candidate generation	2483	2000
3	Item-based collaborative filtering recommendation algorithms	1855	2001
4	Optimizing search engines using click through data	1614	2002
5	The Eigen-trust algorithm for reputation management in P2P networks	1519	2003
6	Mining and summarizing customer reviews	1390	2004
7	Models and issues in data stream systems	1261	2002
8	Maximizing the spread of influence through a social network	1222	2003
9	Data integration: A theoretical perspective	1152	2002
10	What is Twitter, a social network or a news media?	1062	2010
11	Privacy-preserving data mining	998	2000
12	LOF: Identifying density-based local outliers	964	2000
13	A study of smoothing methods for language models applied to Ad Hoc information retrieval	956	2001
14	The skyline operator	914	2001
15	Rdf vocabulary description language 1	873	2004
16	Why we twitter: Understanding microblogging usage and communities	871	2007
17	Earthquake shakes Twitter users: Real-time event detection by social sensors	839	2010
18	GSpan: Graph-based substructure pattern mining	813	2002
19	Generic schema matching with cupid	769	2001
20	Topic-sensitive PageRank	758	2002
21	Relevance based language models	726	2001
22	Yago: A core of semantic knowledge	715	2007
23	Mining the network value of customers	704	2001
24	Pig latin: A not-so-foreign language for data processing	685	2008
25	Training linear SVMs in linear time	684	2006
26	Mining the peanut gallery: Opinion extraction and semantic classification of product reviews	656	2003
27	Group formation in large social networks: Membership, growth, and evolution	631	2006
28	Automatic image annotation and retrieval using cross-media relevance models	631	2003
29	Propagation of trust and distrust	622	2004
30	Rank aggregation methods for the Web	621	2001

pate publishing "The WEKA data mining software: An update", which has the most number of citations among all top conference papers. Thus, both of them have a high number of citations per paper. Kleinberg has published more than 30 papers in these top journals and conferences. However, some of them have received high attention while the others do not. We adopt the average number of citations per paper to measure researchers' impact since it is an accumulative indicator. Here, we assume that if a researcher has published papers and for a long period of time, he/she has more opportunity to receive more citations. As a result, studies [34] usually use c_{10} (the citations received in ten years) to quantify individual's impact avoiding of interference from time accumulation. Beyond that, we also consider the number of publications to refrain from number accumulation because we focus on the impact of each paper.



(a) Journal citation relationship

(b) Conference citation relationship

Figure 6: The overview of Data Mining and Information Retrieval citation relationship among top institutions between 2000 and 2015. Nodes represent institutions and the edges represent citation relationship. (a) The citation relationship among the top 50 most-cited institutions in the top journals. (b) The citation relationship among the top 50 most-cited institutions in the top conferences.

Table 5 shows the top 30 institutions which have the highest average citations per paper in the top journals/conferences. The institutions cluster researchers with essential roles [35]. So the ranking of institutions relies on the researchers' scientific output. The total number of publications and citations of the institution are the sum of researchers' publications and citations who belong to the institution. Most of these institutions have a historical background in Data Mining and Information Retrieval. The geographical distribution of these in-

Table 4: Ranking of authors ordered by the average number of citations per paper between 2000 and 2015.

No.	Author name	Total Citations	Number of Publications	Avg No. of Citations per Paper
1	Peter Reutemann	4594	2	2297
2	Mark A. Hall	4994	3	1664.67
3	Ian H.Witten	4639	3	1546.33
4	Geoffrey Holmes	4994	4	1248.5
5	Eibe Frank	4649	6	774.83
6	Joseph A. Konstan	4274	10	427.4
7	John Riedl	4230	11	384.55
8	Bernhard Pfahringer	4825	13	371.15
9	Erhard Rahm	4549	29	156.86
10	Jon Kleinberg	5203	35	148.66
11	Thorsten Joachims	4314	30	143.8
12	George Karypis	3821	34	112.38
13	Jennifer Widom	6165	66	93.41
14	Philip A. Bernstein	3789	44	86.11
15	Alon Halevy	5181	61	84.93
16	Bing Liu	4307	53	81.26
17	Hector Garciamolina	7256	97	74.80
18	Joseph M.Hellerstein	5174	70	73.91
19	Samuel Madden	5425	78	69.55
20	Jure Leskovec	4208	62	67.87
21	Jian Pei	8056	128	62.94
22	Michael J. Franklin	4901	80	61.26
23	W. Bruce Croft	4085	67	60.97
24	Chengxiang Zhai	5194	88	59.02
25	Eamonn Keogh	4297	73	58.86
26	Jiawei Han	15905	293	54.28
27	Yufei Tao	5004	109	45.91
28	Charu C. Aggarwal	4365	106	41.18
29	Philip S. Yu	9078	259	35.05
30	Christos Faloutsos	5004	155	32.28

stitutions has centered entirely in North America and Asia which means that these two areas have published more papers in the top journals and conferences than others. Most of North America countries belong to the developed countries and have a very high human development index and economic development level. The advanced science and technologies provide the basis for encouraging research results standing in the world. For Asian countries, the research work advances significantly in this era of Data Mining and Information Retrieval.

From the geographical distribution of these institutions, we can find the large differences exist in the citation relationship. In Figure 6, circles represent the top 50 most cited institutions and the size of circles represents the number of self-citations: the larger the circle is, the more self-citations the institution has. The citations obey the distribution in space among journal papers. The top ranking institutions based on the journal papers are centralized in Asia, North America, and Europe primarily. The citation relationships are uniformly distributed throughout these areas. For the conferences papers, citation relationships mainly exist between North America and Asia, Europe and North America. It seems a bit weaker between Europe and Asia. Institutions with higher self-citations exist mainly in North America.

3.4. The inner structure of Data Mining and Information Retrieval

Table 6 present the details of the specific popular keywords for each year. These keywords are the top 1% keywords with the highest occurrence frequency. The research trends focus on “data processing and analysis”, including “information retrieval”, “databases”, “query optimization”, and “query optimization” at the beginning. On the other hand, “machine learning” and “artificial intelligence” are the most investigated topics in the last years.

We recognize different topics based on the method introduced in Section 2. The most popular topics in the first 16 years of 21^{st} Century of journal papers

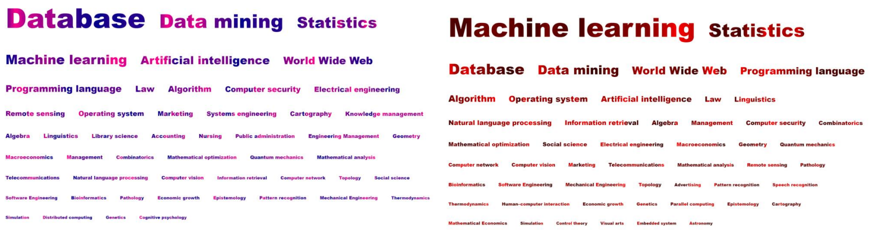
Table 5: Ranking of institutions ordered by the average number of citations per paper between 2000 and 2015.

No.	Institutions	Number of Researchers	Total Number of Citations	Total Number of Publications	Avg No. of Citations per Paper
1	Stanford University	362	54918	847	64.84
2	University of Washington	233	30825	522	59.05
3	Cornell University	229	36640	636	57.61
4	University of Minnesota	188	28899	505	57.23
5	University of California Berkeley	268	28058	558	50.28
6	University of Wisconsin Madison	177	20060	481	41.70
7	University of Massachusetts Amherst	213	23106	596	38.77
8	Massachusetts Institute of Technology	228	16544	439	37.69
9	Carnegie Mellon University	437	38601	1047	36.87
10	University of Illinois at Urbana Champaign	338	36998	1140	32.45
11	IBM	1409	96866	3316	29.21
12	Purdue University	212	17817	611	29.16
13	Hong Kong University of Science and Technology	218	21189	807	26.26
14	Yahoo	733	49364	1923	25.67
15	Microsoft	1288	93165	3728	24.99
16	Google	513	18936	832	22.76
17	University of Maryland College Park	255	12595	589	21.38
18	University of Waterloo	164	9687	462	20.97
19	Max Planck Society	135	9516	463	20.55
20	Arizona State University	183	9609	491	19.57
21	National University of Singapore	407	21353	1129	18.91
22	The Chinese University of Hong Kong	207	12990	696	18.66
23	Pennsylvania State University	172	7193	431	16.69
24	Tsinghua University	372	13479	981	13.74
25	Oracle Corporation	241	5273	484	10.89
26	Nanyang Technological University	174	5327	491	10.85
27	Peking University	267	5229	497	10.52
28	National Taiwan University	205	4561	449	10.16
29	Zhejiang University	179	3806	468	8.13
30	Chinese Academy of Sciences	307	4537	562	8.07

Table 6: Ranking of popular keywords ordered by the occurrence probability between 2000 and 2015.

Year	Rate	Hot Keywords
2000	0.46	information retrieval, data mining, world wide web, indexation, internet, databases, information technology, information science, image retrieval, information system, database system, computer science, query language
2001	0.47	data mining, indexation, information retrieval, computer science, query optimization, web pages, indexing, database system, search engine, learning artificial intelligence, world wide web, data analysis, association rules, xml, databases, xml document
2002	0.51	data mining, information retrieval, computer science, indexation, search engine, relational databases, xml, web pages, association rules, dynamic content, data engineering, internet, query optimization, performance, satisfiability, information management, databases
2003	0.48	data mining, information retrieval, indexation, information system, xml document, web service, web pages, search engine, semantic web, classification, relational databases, xml, learning artificial intelligence, geographic information systems, computational complexity, information management
2004	0.53	data mining, information retrieval, indexation, web pages, semantic web, indexing terms, search engine, xml, query optimization, web service, internet, tree data structures, xml document, clustering, information system, satisfiability, computational complexity
2005	0.59	data mining, information retrieval, indexation, indexing terms, web pages, xml, computer science, databases, relational databases, internet, search engine, data engineering, semantic web, xml document, web service, learning artificial intelligence, data analysis, information system, query optimization, information technology, database system
2006	0.57	data mining, information retrieval, indexation, computer science, xml, data engineering, web pages, databases, information system, learning artificial intelligence, machine learning, information technology, search engine, information management, data analysis, geographic information systems, information security, indexing terms, information science, management information systems, satisfiability, semantic web, business model, soft system methodology
2007	0.52	data mining, information retrieval, indexation, computer science, xml, search engine, information system, databases, information management, web pages, machine learning, information technology, information security, internet, information science, geographic information systems, data analysis, relational databases, management information systems, business model, information management system, soft system methodology, computer information systems, accounting information systems, information systems technology
2008	0.59	data mining, databases, information retrieval, indexation, computer science, algorithm design and analysis, data models, probability density function, xml, search engine, clustering algorithms, indium, data analysis, web pages, arsenic, internet, indexes, learning artificial intelligence, machine learning, feature extraction, information system, accuracy, beryllium, optimization, relational databases, semantic web, satisfiability, information technology
2009	0.60	data mining, databases, probability density function, information retrieval, indexation, search engine, learning artificial intelligence, clustering algorithms, data models, machine learning, algorithm design and analysis, web pages, information system, accuracy, internet, data analysis, satisfiability, social network, database system, information management, information technology, indexes, clustering, business model, optimization
2010	0.61	data mining, information retrieval, databases, indexation, search engine, learning artificial intelligence, data models, social network, clustering algorithms, machine learning, computer science, algorithm design and analysis, data analysis, internet, optimization, accuracy, web pages, satisfiability, feature extraction, semantics, indexing, graph theory, xml, relational databases, computational modeling, indexes, database management systems, uncertainty, information system
2011	0.60	data mining, indexation, information retrieval, search engine, social network, learning artificial intelligence, semantics, graph theory, databases, machine learning, internet, web pages, data analysis, satisfiability, data handling, recommender system, algorithm design and analysis, data model, information system, algorithm design, social media, text analysis, database system, indexes, xml
2012	0.52	data mining, social network, information retrieval, indexation, learning artificial intelligence, search engine, graph theory, machine learning, social media, databases, data analysis, algorithm design and analysis, internet, data models, database management systems, information system, geographic information systems, information management, indexing, real time, social networks, business model, information management system
2013	0.38	data mining, learning artificial intelligence, graph theory, information retrieval, data analysis, data models, semantics, indexes, social networks, social media, databases, recommender systems, data handling, vectors, internet, optimization, algorithm design and analysis, indexing
2014	0.30	data mining, vectors, data models, algorithm design and analysis, accuracy, optimization, databases, semantics, clustering algorithms, feature extraction, indexes, computational modeling, measurement, mathematical model
2015	0.28	data mining, data models, algorithm design and analysis, indexes, optimization, computational modeling, approximation algorithms, semantics, databases, geographic information systems, information management, clustering algorithms, information technology, business model

are “Database”, “Data Mining”, and “Statistics” (see in Figure 7(a)). Meanwhile, popular topics of conference papers are similar to journal papers. The popular topics of conference papers are “Machine Learning”, “Statistics”, and “Database” (see in Figure 7(b)). The size of each topic reflects its overall popularity rankings during 2000-2015, and the color variation of each topic shows the change of its popularity within 2000-2015 on the yearly basis. The brighter the color is, the more popular the topic is. For example, in Figure 7(b), we can see that “Machine Learning” is getting more popular and “Database” is getting less popular over time.



(a) Popular topics for journal papers

(b) Popular topics for conference papers

Figure 7: The evolution of topics over time. The change in color from left to right reflects the annual heat of the topic. Sizes of the topics reflect the rankings in 2000-2015.

Furthermore, for each of the popular topics, we draw a word cloud of the topics associated with the most relevant topics (see in 5). Examples of other popular topics during this time period include: “Data Mining”, “Statics”, “World Wide Web”, and “Artificial Intelligence”. It also shows that Data Mining and Information Retrieval specializes theories, technologies and methods for various research fields. As for specialized Data Mining and Information Retrieval, many sub-fields have developed dramatically , i.e., Behavioral Data Science, Life Data Science, Brain Data Science, Meteorology Data Science, Financial Data Science, Geography Data Science, and so on.

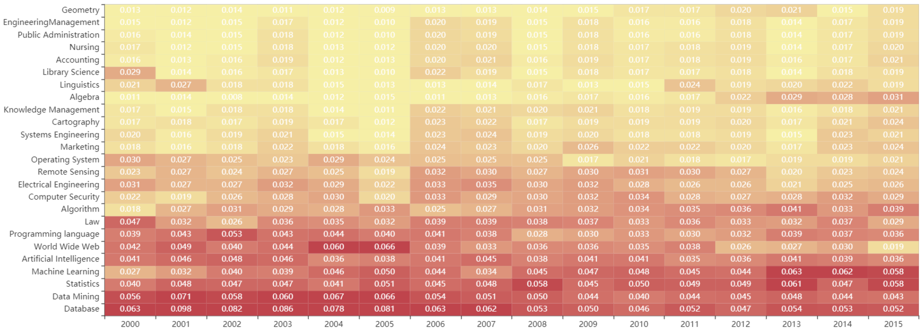
We have discussed the popular keywords and topics for the study period above. In order to give a deeper insight into the evolution of topics, we unveil the evolutionary pattern via calculating development index DI each year (Figure 8) and increasing index II (Table 7) for these topics. Both of them can reflect the temporal trend of the topic. Specifically, DI illustrates the proportion of the topic over time. The popularity of some topics have been increasing (i.e., Statics), while others become weak (i.e., Data Mining), and fluctuated (i.e., Quantum Mechanics). The distinction between topics in journal articles and those in conference papers can also be found easily through in Figure 8.

Table 7: Increase index for popular topics.

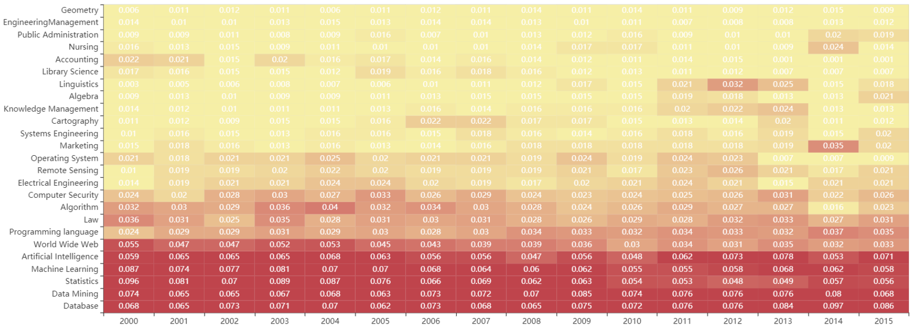
Topic	II	Topic	II	Topic	II	Topic	II	Topic	II
Advertising	2.67	Economic Growth	1.28	Knowledge Management	1.06	World Wide Web	0.90	Systems Engineering	0.78
Social Science	2.12	Human-computer Interaction	1.26	Linguistics	1.01	Law	0.90	Topology	0.75
Mathematical Optimization	1.53	Algorithm	1.20	Computer Security	1.00	Cartography	0.88	Electrical Engineering	0.75
Mathematical Analysis	1.52	Parallel Computing	1.20	Management	1.00	Remote Sensing	0.88	Programming Language	0.75
Algebra	1.46	Computer Vision	1.19	Nursing	1.00	Computer Network	0.85	Database	0.74
Quantum Mechanics	1.44	Machine Learning	1.17	Telecommunications	0.99	Genetics	0.83	Bioinformatics	0.73
Mechanical Engineering	1.36	Natural Language Processing	1.13	Speech Recognition	0.97	Data mining	0.83	Library science	0.73
Thermodynamics	1.35	Statistics	1.12	Accounting	0.94	Information retrieval	0.80	Pattern Recognition	0.64
Combinatorics	1.32	Marketing	1.12	Public Administration	0.94	Epistemology	0.80	Software Engineering	0.59
Geometry	1.32	Pathology	1.09	Operating system	0.93	Artificial Intelligence	0.79	Macroeconomics	0.52

As DI reports the changes in the importance of these topics over time, II can summarize the overall trend of popularity. $II > 1$ indicates that topics become more popular in 2008-2015 than 2000-2007, it is applicable in reverse. The result provides a base for further study in topics evolution.

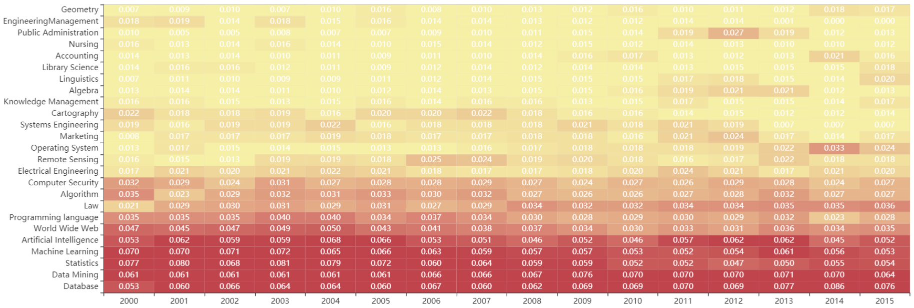
If we use Figure 8 and Table 7 together to analyze the evolution of popular topics, although the proportion varies over time, the order of popular topics



(a) Journals



(b) Conferences



(c) All

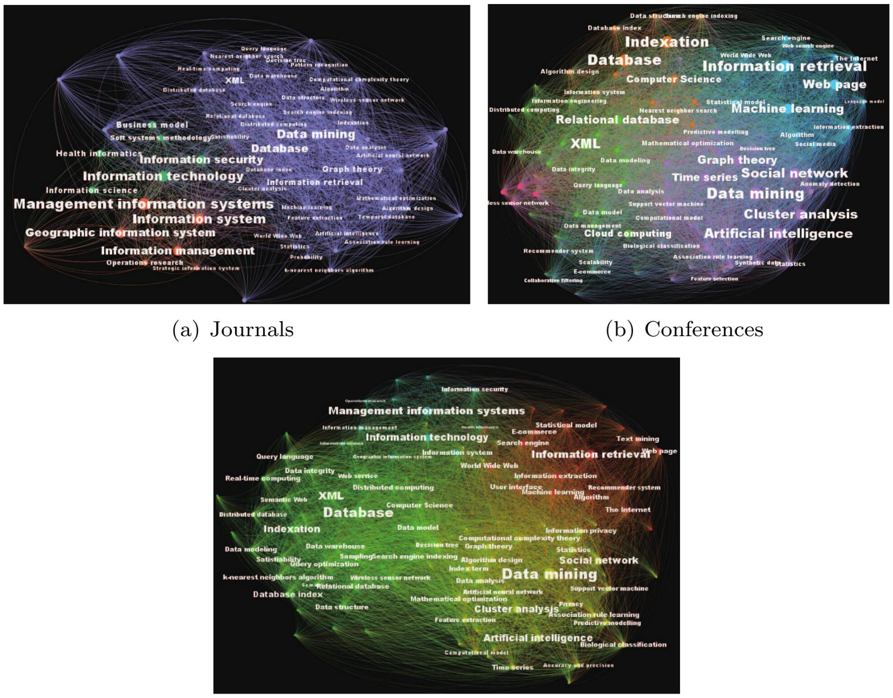
Figure 8: The evolution of topics over time.

remains unchanged. That is, the focuses have stabilized over the study period. For example, although the topic “Database” becomes less popular in 2008-2015, publications on this topic always account for a large proportion. On the other hand, the topic “Mechanical Engineering” becomes popular comparing with the first 8 years, buy it is still not popular enough evidenced by its low DI. The overall trend of topic development needs to be analyzed in conjunction with DI and II.

From the analysis of topics distribution, we can find strong interconnections among some topics. The citation relationship among these topics can reveal the correlation in different topics. To measure the conceptual distance between topics, we visualize the network of the citations relationship across all topics (Figure 9). Nodes in this network represent specific topics and edges are captured from the citation relationship in papers. Clusters of topics (represented by different colors) which are highly connected are clearly shown in this figure. Meanwhile, the large differences in the journals’ topics and conferences’ topics can be easily found. For example, the topic “Data Mining” appears frequently in the topic citation network and closely connect with similarly topics such as “Data Analysis”, “Data Structure”, and “Graph Theory” (see the clusters in purple of Figure 9(a)). Meanwhile, for conferences’ topics, the topics can be divided into more clusters on behalf of “Information Retrieval”, “Databases”, “Data Mining” and “Cloud Computing” (Figure 9(a)). In general, the network presents how far a pair of topics are from each other and how different these topics are allocated in the era of Data Mining and Information Retrieval. It can be used as a tool to detect topics with diverse representations in different sub-fields of Data Mining and Information Retrieval.

4. Conclusion

In this paper, we provide an empirical analysis to discover the anatomy of Data Mining and Information Retrieval spanning the first 16 years of the 21^{st} Century. We aim at providing evolutionary and dynamically changing views of



(a) Journals

(b) Conferences

(c) All

Figure 9: The citation relationships among topics based on journals and conferences. A node represents a topic and its size based on the number of publications. Different colors represent different clusters of sub-fields.

its development. From the publication metadata, we find the gradual, continual process of steady growth of productivity, impact, and collaboration concerning the number of publications, citations, and the average number of authors per paper, respectively. From the perspective of reference behavior, the age of the literature that researchers focus on has been increasing from 2000 to 2009. After that, the new literature has become the state of the art. Nearly 70%-80% declination in the self-citation rate illustrates that the area is becoming more open-minded and sharing. Most influential papers/researchers/institutions are identified as the leaders of innovation in Data Mining and Information Retrieval. Finally, by exploring the inner structure, the dynamic in topics reveals the development of this field: some topics have remained consistent over the study period, e.g., “Data mining”, “Information retrieval”, and “Databases”. In contrast, other topics have been bereft prevalence. These results help researchers understand the research trends and emphasize their research focuses. The citation relationship among these topics are clustered into different groups, which can be used as a tool to measure conceptual distance between topics.

In summary, our study reveals the nature and evolution of the Data Mining and Information Retrieval field and identifies the main insights behind the existing knowledge. However, the generalisability of these results is subject to certain limitations. For instance, in-depth techniques such as co-word and co-citation analysis could be used to recognize the diversity of sub-fields and find the pattern of how do they interact with each other. In the future, we will focus more on the network structure and use network-based metrics to evaluate the dynamic of this interdisciplinary area. Another trend of potential research is adopting other probabilistic models, such as Latent Dirichlet Allocation (L-DA) [36] to identify dynamic topic patterns and discover more insights in large corpus of Data Mining and Information Retrieval literature.

5. Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61872054), and the Fundamental Research Funds for the Central Universities (DUT19LAB23, DUT18JC09).

References

[1] D. J. Hand, Data mining, Encyclopedia of Environmetrics 2.

[2] H. M. Chung, F. Gey, S. Piramuthu, Data mining and information retrieval, in: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, IEEE, 2002, pp. 841–842.

[3] F. Xia, W. Wang, T. M. Bekele, H. Liu, Big scholarly data: A survey, IEEE Transactions on Big Data 3 (1) (2017) 18–35.

[4] S. Khan, X. Liu, K. A. Shakil, M. Alam, A survey on scholarly data: From big data perspective, Information Processing & Management 53 (4) (2017) 923–944.

[5] D. Gu, J. Li, X. Li, C. Liang, Visualizing the knowledge structure and evolution of big data research in healthcare informatics, International journal of medical informatics 98 (2017) 22–32.

[6] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, H. E. Stanley, The science of science: From the perspective of complex systems, Physics Reports 714-715 (16) (2017) 1–73.

[7] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al., Science of science, Science 359 (6379) (2018) eaao0185.

[8] G. Abramo, C. A. DAngelo, G. Murgia, Gender differences in research collaboration, Journal of Informetrics 7 (4) (2013) 811–822.

[9] J. Tang, S. Wu, J. Sun, H. Su, Cross-domain collaboration recommendation, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1285–1293.

[10] W. Wang, X. Bai, F. Xia, T. M. Bekele, X. Su, A. Tolba, From triadic closure to conference closure: The role of academic conferences in promoting scientific collaborations, Scientometrics 113 (1) (2017) 177–193.

[11] W. Wang, J. Liu, F. Xia, I. King, H. Tong, Shifu: Deep learning based advisor-advisee relationship mining in scholarly big data, in: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2017, pp. 303–310.

[12] J. Jiang, P. Shi, B. An, J. Yu, C. Wang, Measuring the social influences of scientist groups based on multiple types of collaboration relations, Information Processing & Management 53 (1) (2017) 1–20.

[13] X. Kong, M. Mao, W. Wang, J. Liu, B. Xu, Voprec: Vector representation learning of papers with text information and structural identity for recommendation, IEEE Transactions on Emerging Topics in Computing.

[14] Y. Sun, J. Han, Mining heterogeneous information networks: Principles and methodologies, Synthesis Lectures on Data Mining and Knowledge Discovery 3 (2) (2012) 1–159.

[15] W. Wang, S. Yu, T. M. Bekele, X. Kong, F. Xia, Scientific collaboration patterns vary with scholars' academic ages, Scientometrics 112 (1) (2017) 329–343.

[16] L. Sun, I. Rahwan, Coauthorship network in transportation research, Transportation Research Part A: Policy and Practice 100 (2017) 135–151.

[17] M. Meyer, I. Lorscheid, K. G. Troitzsch, The development of social simulation as reflected in the first ten years of JASSS: A citation and co-citation analysis, Journal of Artificial Societies and Social Simulation 12 (4) (2009) 12.

[18] R. Sinatra, P. Deville, M. Szell, D. Wang, A.-L. Barabási, A century of physics, Nature Physics 11 (10) (2015) 791–796.

[19] W. Iqbal, J. Qadir, G. Tyson, A. N. Mian, S.-u. Hassan, J. Crowcroft, A bibliometric analysis of publications in computer networking research, Scientometrics 119 (2) (2019) 1121–1155.

[20] I. Lee, F. Xia, G. Roos, An observation of research complexity in top universities based on research publications, in: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2017, pp. 1259–1265.

[21] T. Amjad, Y. Ding, J. Xu, C. Zhang, A. Daud, J. Tang, M. Song, Standing on the shoulders of giants, Journal of Informetrics 11 (1) (2017) 307–323.

[22] Y. Dong, H. Ma, Z. Shen, K. Wang, A century of science: Globalization of scientific collaborations, citations, and innovations, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 1437–1446.

[23] A. Firdaus, M. F. Ab Razak, A. Feizollah, I. A. T. Hashem, M. Hazim, N. B. Anuar, The rise of blockchain: bibliometric analysis of blockchain study, Scientometrics (2019) 1–43.

[24] J. Liu, X. Kong, F. Xia, X. Bai, L. Wang, Q. Qing, I. Lee, Artificial intelligence in the 21st century, IEEE Access 6 (2018) 34403–34421.

[25] P. Palvia, C. P. YK, M. D. Kakhki, T. Ghoshal, V. Uppala, W. Wang, A decade plus long introspection of research published in Information & Management, Information & Management 54 (2) (2017) 218–227.

[26] M. R. Frank, D. Wang, M. Cebrian, I. Rahwan, The evolution of citation graphs in artificial intelligence research, Nature Machine Intelligence 1 (2) (2019) 79.

[27] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, K. Wang, An overview of microsoft academic service (MAS) and applications, in: Proceedings of the 24th International Conference on World Wide Web.

[28] D. Aksnes, A macro study of self-citation, Scientometrics 56 (2) (2003) 235–246.

[29] K. Hyland, Self-citation and self-reference: Credibility and promotion in academic publication, Journal of the Association for Information Science and Technology 54 (3) (2003) 251–259.

[30] F. Radicchi, C. Castellano, Rescaling citations of publications in physics, Physical Review E 83 (4) (2011) 046116.

[31] B.-C. Björk, D. Solomon, The publishing delay in scholarly peer-reviewed journals, Journal of Informetrics 7 (4) (2013) 914–923.

[32] A. Correia, H. Paredes, B. Fonseca, Scientometric analysis of scientific publications in CSCW, Scientometrics 114 (1) (2018) 31–89.

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: An update, ACM SIGKDD explorations newsletter 11 (1) (2009) 10–18.

[34] R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási, Quantifying the evolution of individual scientific impact, Science 354 (6312) (2016) aaf5239.

[35] O. Mubin, A. Al Mahmud, M. Ahmad, Hci down under: Reflecting on a decade of the OzCHI conference, Scientometrics 112 (1) (2017) 367–382.

[36] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (Jan) (2003) 993–1022.

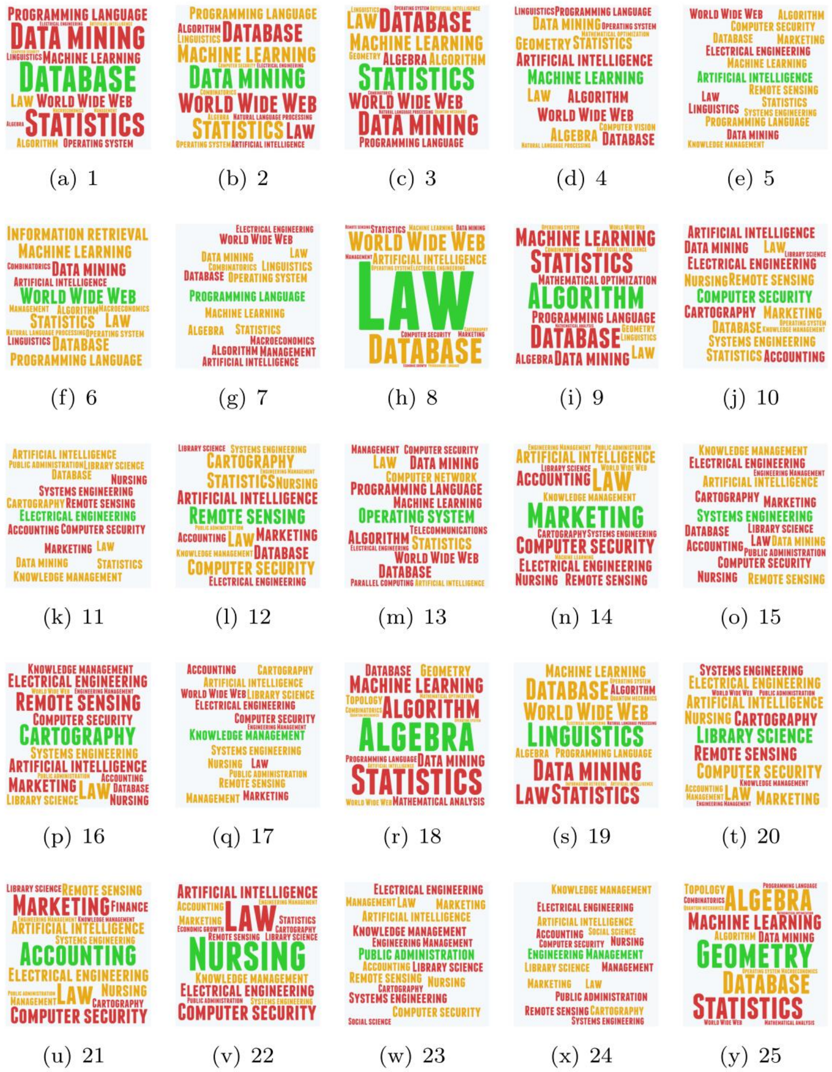
Supplemental files: Word clouds for popular topics

The relevance r between topics is calculated based on the occurrence of two topics. It is can be computed as:

$$r = (n_{ab}/n)/(n_b/n) = n_{ab}/n_b \tag{Eq. (A.1)}$$

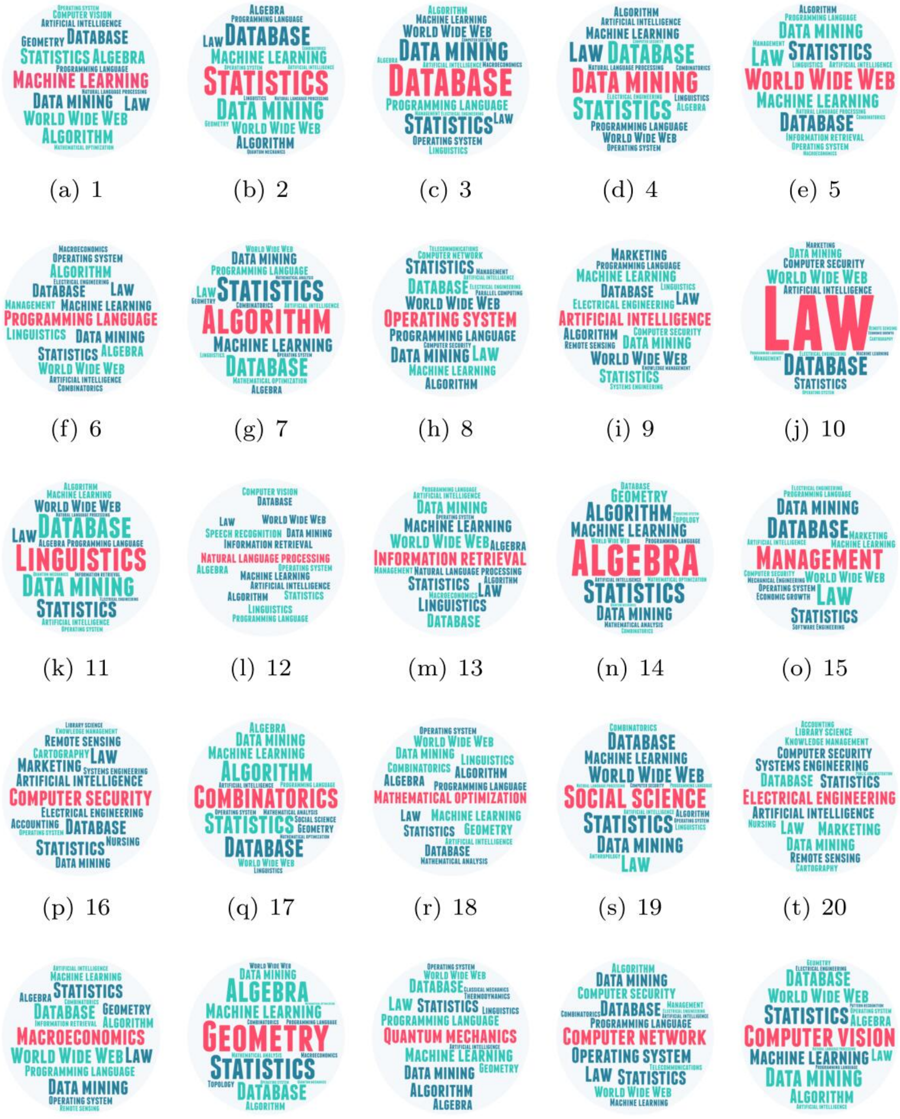
where n is the total number of publications, n_{ab} represents the number of publications which belong to topic a and b simultaneously, n_b is the number of publications which belong to the topic b.

Word clouds of the popular topics and their relevant topics are present in Figure A.1 and Figure A.2.



(a) 1
(b) 2
(c) 3
(d) 4
(e) 5
(f) 6
(g) 7
(h) 8
(i) 9
(j) 10
(k) 11
(l) 12
(m) 13
(n) 14
(o) 15
(p) 16
(q) 17
(r) 18
(s) 19
(t) 20
(u) 21
(v) 22
(w) 23
(x) 24
(y) 25

Figure A.1: The word cloud based on the topic of journals in high occurrence probability in the 21^{st} Century.



(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

(f) 6 (g) 7 (h) 8 (i) 9 (j) 10

(k) 11 (l) 12 (m) 13 (n) 14 (o) 15

(p) 16 (q) 17 (r) 18 (s) 19 (t) 20

(u) 21 (v) 22 (w) 23 (x) 24 (y) 25

Figure A.2: The word cloud based on the topic of conferences in high occurrence probability in the 21^{st} Century.